La r-confiance pour l'identification de trajectoires de patients

Yves Mercadier, Jessica Pinaire, Jérôme Azé, Sandra Bringay, Maguelonne Teisseire

La r-confiance pour l'identification de trajectoires de patients.

Yves Mercadier, Jessica Pinaire, Jérôme Azé, Sandra Bringay, Maguelonne Teisseire

LIRMM, UMR 5506, Université Montpellier, France

Objectifs

La phase d'extraction des motifs en fouille de données exploratoire conduit souvent à générer un trop gros volume de motifs à valider [1].
L'expert est alors submergé et démuni devant les nouvelles données ainsi produites. Notre proposition d'une nouvelle mesure de filtrage s'inscrit dans ce contexte et a été implémentée dans une interface interactive avec pour objectif de faciliter l'analyse des résultats de la fouille de données séquentielles.

Introduction

Nous proposons une mesure originale, appelée **r-confiance**, qui présente un double intérêt : (1) elle fonctionne pour tous les types de motifs (règle d'association, motif séquentiel, motif spatio-temporel) et (2) elle utilise comme opérateur d'agrégation «la proportion de position».

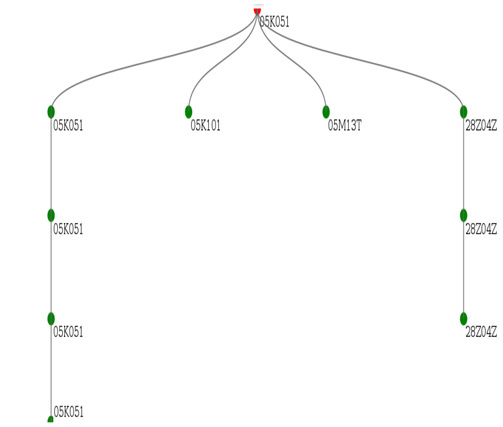

Figure 1: Représentation sous forme d'arbre de l'ensemble des motifs débutants par l'item "infarctus du myocarde".

Nous avons également développé une interface qui permet de représenter les ensembles de motifs extraits mais également de les comparer en prenant en compte différentes mesures d'intérêt [2], dont la r-confiance (L'interface sera présentée lors de la session démonstration).

r-confiance élémentaire

Avant de définir la r-confiance d'un motif de façon générale, nous définissons la r-confiance élémentaire d'un motif séquentiel.
Étant donné une base de données B et un motif séquentiel $M =< M_1, M_2, ..., M_n >$, un **candidat séquentiel** de M, $C =< M_1, M_2, ..., M_p >$, est défini comme une des sous-séquences préfixes de p items de M telle que $p < n$.
Un motif séquentiel de longueur n est ainsi associé à $n - 1$ candidats séquentiels.

Résultat

$$r\text{-}conf(M) = \begin{cases} 0 & \text{si } Card(\{C \in \mathbf{C}, r\text{-}conf\text{-}e(M,C) > minR\}) = 0 \\ \frac{Card(\{C \in \mathbf{C}, r\text{-}conf\text{-}e(M,C) > minR\})+1}{n} & \text{sinon} \end{cases} \quad (1)$$

Soit M un motif et C un candidat séquentiel de ce motif. La r-confiance élémentaire, notée *r-conf-e*, est définie à partir des supports des séquences impliquées par :

Définition

$$r\text{-}conf\text{-}e(M,C) = \frac{support_B(M)}{support_B(C)} \quad (2)$$

La r-confiance calculée pour le motif M correspond à l'agrégation des $n - 1$ r-confiances élémentaires des candidats séquentiels le composant.
Afin de conserver la notion de mesure d'intérêt et donc de filtrage des motifs extraits, seules les r-confiances élémentaires dont la valeur est supérieure à un seuil fixé $minR$ seront prises en compte dans cette agrégation.

Application à l'identification de parcours hospitaliers

Nous avons étudié les parcours de soins des patients atteints d'un Infarctus du Myocarde au cours de la période 2009-2013. Ces données sont issues des bases hospitalières nationales du PMSI (Programme de Médicalisation du Système d'Information). L'évaluation de cette mesure par la mise en évidence de parcours connus du corpus médical est encourageante. Par la suite, nous souhaitons appliquer cette mesure dans l'identification de parcours hospitaliers types [3].
Ces résultats prometteurs vont permettre l'identification de trajectoires de patients et de poursuivre les travaux d'exploration des parcours.

Résultat

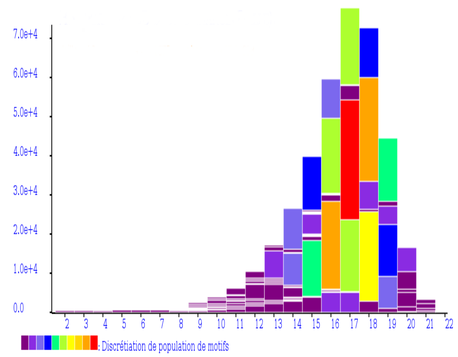

Figure 2: Histogramme empilé de répartition de r-confiance par rapport à la taille des motifs

Conclusion

Nous avons proposé une nouvelle mesure d'intérêt qui est une extension de la confiance, définie pour les règles d'association, aux motifs séquentiels. Un expert en cardiologie est actuellement sollicité pour évaluer l'impact de la mesure proposée dans la validation des connaissances extraites, réel objectif d'une telle étude.

Références

[1] Rakesh Agrawal, Tomasz Imielinski, and Arun Swami. Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, pages 207–216, New York, NY, USA, 1993.

[2] Julien Blanchard. Un système de visualisation pour l'extraction, l'évaluation, et l'exploration interactives des règles d'association., 2005.

[3] Anders Boeck Jensen, Pope Moseley, Tudor Oprea, Sabrina Gade Ellesøe, Robert Eriksson, Henriette Schmock, Peter Bjødstrup Jensen, Lars Juhl Jensen, and Søren Brunak. Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients. volume 5 of *Nature Communications*, page 4022, 2014.

Informations de Contact

- Web: `http://www.lirmm.fr/`
- Email: yves.mercadier@ac-montpellier.fr